AMNet: Memorability Estimation with Attention

Jiri Fajtl[1], Vasileios Argyriou[1], Dorothy Monekosso[2], Paolo Remagnino[1]
[1]Kingston University, London, UK
[2]Leeds Beckett University, Leeds, UK

Abstract

In this paper we present the design and evaluation of an end-to-end trainable, deep neural network with a visual attention mechanism for memorability estimation in still images. We analyze the suitability of transfer learning of deep models from image classification to the memorability task. Further on we study the impact of the attention mechanism on the memorability estimation and evaluate our network on the SUN Memorability and the LaMem datasets. Our network outperforms the existing state of the art models on both datasets in terms of the Spearman's rank correlation as well as the mean squared error, closely matching human consistency.

1. Introduction

The ability of man cognition to recall as well as forget visual content after viewing it is very important to the way we acquire new information and interact with our environment. This is becoming increasingly significant as creating and consuming visual content dominates other forms of information exchange. Moreover, low cost, automated image and video capture systems are rapidly surfacing as the norm in the Internet of the Things (IoT) domain, also contributing to the visual information flow.

To which degree an image is later remembered or forgotten is expressed as image memorability. It is an important cognitive measure to be taken into account while processing visual content, whether for human to human or machine to human communication or for storage.

Memorability estimation has a large variety of practical applications, such as selecting or designing highly memorable advertising material, organizing and tagging of photos in albums, introducing a real-time, image memorability measure built into consumer digital cameras, helping to make highly memorable presentations and data visualizations, improving memorability of specific parts of a graphical user interface (GUI) or helping to illustrate education material. An application of a great interest is to measure a decline in memory capacity of patients affected by dementia (such as Alzheimer's and Parkinson's diseases) and forms of mild cognitive impairment (MCI).

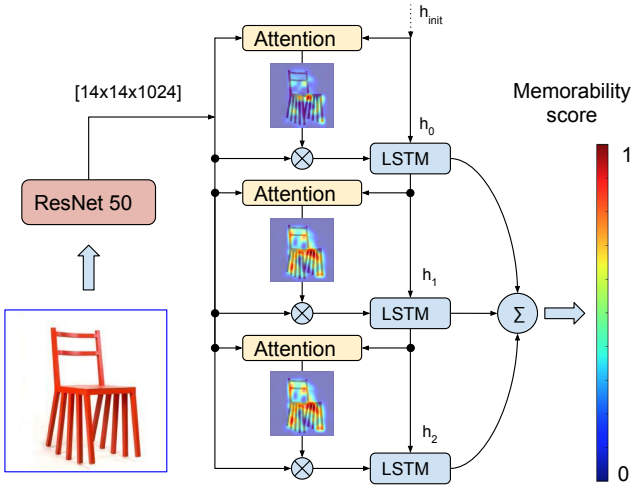

Figure 1: AMNet iteratively generates attention maps linked to the image regions correlated with the memorability. After three iterations the memorability scores are added and presented on the output.

Prior research [13] has shown that image memorability has a stable property, that is, individuals tend to remember the same images with the same probability regardless of delays, and that it can be quantified and measured. This research has led to first attempts to learn and predict memorability with machine learning freameworks, initially with low-level, global image features [12], reaching moderate success. To improve such a solution would, however, require the design of new features, which demands a strong domain knowledge not well understood in the specific case of memorability.

In [37], [2] and recently [14] has been shown that this problem can be mitigated by applying deep learning techniques to the memorability domain. Deep learning, however, requires large training dataset which was not available until A. Khosla et al. [18] introduced a large memorability dataset LaMem with 60K images and subsequently used it

to train the MemNet, which is based on the AlexNet [21] initialized on the ImageNet [30] and Places [39] datasets. MemNet achieves Spearman's rank correlation $\rho = 0.64$ compared with the human consistency $\rho = 0.68$ as measured by [18].

Intuitively, image regions immediately drawing our attention would appear to be linked with highly memorable visual content. Indeed, this assumption was confirmed to be correct in the works of [24], [19] and [13] who already very early indicated a potential relationship between the visual attention and memorability but did not further investigate their correlation. To that end, we propose the Attention based Memorability estimation Network-AMNet, a novel, deep neural network architecture with a recurrent, visual attention mechanism with the primary goal to improve on the state of the art for the memorability prediction task. We also show advantages of the visualization of the generated attention maps and their connection to the memorability property. Our approach is extensively evaluated on the LaMem [18] and SUN Memorability [13] datasets. The main contributions of our work are:

- AMNet as a generic architecture for regression tasks with deep CNN, visual attention mechanism and recurrent neural network.
- application of the proposed AMNet to the image memorability estimation.
- introduction of the incremental memorability estimation with the recurrent network and demonstration of the achieved performance gain.
- introduction of the visual attention technique for the memorability estimation and presentation of the performance gain.
- demonstration that transfer learning from deep models, trained for image classification, is particularly beneficial for the memorability estimation.

The paper is organized as follows: Section 2 provides background material on image memorability, its properties, measurement and prediction. In section 3 we propose the AMNet and discuss the theoretical framework behind this architecture and the training procedure. The performance of AMNet is studied in the section 4, with section 5 concluding this work.

2. Previous Work

In a pioneering work on image memorability, Isola et al. [13], [11] demonstrated that the ability of our cognition system to remember certain images and forget other is congruent among independent observers, despite large variability in the image content, concluding memorability is a stable property, intrinsic to images. Based on this premise, Isola et al. [13] investigated factors that give rise to the image memorability effect, which was then used to predict image memorability scores with a machine learning program, based on global image features GIST [26], SIFT [22], HOG [5], SSIM [33] and pixel histogram.

In order to build better computational models to learn and predict memorability, researchers analyzed the relationship between memorability and various visual factors [19], image classes [12] and saliency [7]. Bylinskii et al. [3] conducted a number of experiments to better understand the intrinsic and extrinsic effects on image memorability, concluding that the primary substrate of memorability lies in the intrinsic properties of images and all extrinsic effects contribute only marginally.

Deep learning was first applied to the memorability problem by Baveye et al. [2] who proposed a MemoNet model based on GoogLeNet [34] trained on the ImageNet [30] dataset. [37] used CNN features with SVR [6] to predict memorability with accuracy comparable to MemoNet [2].

To achieve higher accuracy with deep learning techniques Khosla et al. [18] collected a large memorability dataset LaMem with 60K images and introduced MemNet model based on the Hybrid-CNN, which is the AlexNet [21] CNN pretrained on the ImageNet [30] and the Places [39] datasets ($\sim$3.6 million images in total). Researchers also tried to improve memorability prediction by other techniques, such as the adaptive transfer learning from external sources [14] or predicting image memorability by multi-view adaptive regression [27], none exceeding the performance of the MemNet [18].

Relationship between the visual attention and memorability was already suggested by Isola et al. [13] but was not further investigated. Mancas and Le Meur [24] studied the link between saliency and memorability and found that the most memorable images have uniquely localized regions, while less memorable either do not have precise regions of interest or have several of them. Based on these findings, [24] devised new attention-related features that improved the memorability prediction by 2% compared to the non attention based models from [13]. In a similar work, Celikkale et al. [4] applied an attention driven spatial pooling pipeline based on SIFT [22] and HOG [5] features and bottom-up and object-level saliency detectors. Their results, albeit only moderate, still indicate a benefit of the attention based approach. Importance of the memorability regions was explored by Khosla et al. [19] who introduced the concept of attention maps that relate image regions to memorability. These maps are learnt directly as clusters of gradients, textures and color features with the SVM-Rank solver [15] with results showing benefits of the attention on memorability prediction.

In our work we investigate the application of deep learn-

ing methods with visual attention and recurrent network to learn and predict image memorability. To our knowledge the presented approach has not been attempted before.

3. Method

The idea behind the AMNet architecture is based on four main components a deep CNN trained on large-scale image classification task, a soft attention network, a Long Short Term Memory (LSTM) [9] recurrent neural network followed by a fully connected neural network for memorability score regression.

In the following section we introduce the details of the AMNet architecture as shown in Figure 2, starting with the pre-trained CNN (a) for transfer learning. Subsequently we show the working of the visual, soft attention mechanism (b), the LSTM and network for the memorability regression (c) and (d). Finally we outline the training procedure and finish with the data augmentation process.

3.1. Transfer Learning for Memorability Estimation

It is common practice to use a pretrained CNN as a fixed feature extractor or to fine tune it for a similar application [32], mainly to reduce training time and overfitting on tasks with small datasets.

This technique is readily applied to computer vision problems centered around semantic features such as objects detection and segmentation, however little is known about such transfer learning for the image memorability estimation since there is no clear understanding of what visual features trigger the effects of remembering and forgetting.

Khosla et al. [18] has already shown the benefits of fine tuning of pretrained CNN for this domain, however we decided to evaluate a much deeper model as a fixed feature extractor. Our results show that the features learnt for image classification are highly suitable for the memorability task. In our work we use ResNet50 [10] model trained on ImageNet where it achieves the top 1 error 24.7%.

3.2. Soft Attention Mechanism

The ability of a neural network to learn which discrete information elements to focus on within a given training sample was first applied in machine translation by Bahdanau et al. [1]. This mechanism is called soft attention due to the fact that it produces a probability weight for every information element rather than a hard decision boundary. The benefit of soft attention is that it can be learnt end-to-end

with a gradient based optimization method.

The soft attention mechanism has two components, a network that learns probabilities for each information element within the input data and a gating function that uses these probabilities to weigh data for further processing.

3.3. AMNet Details

The AMNet estimates the image memorability by taking a single image $\boldsymbol{X}$ and generating a memorability score y.

$$y = f(\boldsymbol{X}), \quad y = [0, 1] \tag{1}$$

The process of memorability estimation is summarized in algorithm 1.

Algorithm 1 AMNet algorithm

procedure MEMORABILITY(X) ▷ $y = f(\boldsymbol{X})$
  $\mathbf{x} = \textit{get\_cnn\_features}(\boldsymbol{X})$ ▷ ResNet50 fwd pass
  $\mathbf{h}_0 = f_{init_c}(\mathbf{x})$ ▷ Eq. 12
  $\mathbf{c}_0 = f_{init_h}(\mathbf{x})$ ▷ Eq. 12
  $lstm\_init(\mathbf{h}_0, \mathbf{c}_0)$
  $y = 0$
  for $t = 0$ to T **do** ▷ at $t = 0 \rightarrow \mathbf{h}_t = \mathbf{h}_0$
    $\mathbf{e} = f_{att}(\mathbf{x}, \mathbf{h}_t)$ ▷ Eq. 8
    $\boldsymbol{\alpha} = softmax(\mathbf{e})$ ▷ Eq. 6
    $\mathbf{z} = []$
    for $i = 0$ to L **do** ▷ for all locations, Eq. 4
      $\mathbf{z} = \mathbf{z} + \alpha_i \mathbf{x}_i$ ▷ $\mathbf{z} \in \mathbb{R}^D$
    $\mathbf{h}_t, \mathbf{c}_t = lstm\_step(\mathbf{z}, \mathbf{h}_t, \mathbf{c}_t)$ ▷ Eq. 3
    $y = y + f_m(\mathbf{h}_t)$ ▷ Eq. 11
  return y ▷ Memorability score $[0, 1]$

Formally, let the image features, extracted by a CNN, be a tensor with dimensions (W, H, D) where W and H represent the spatial resolution while D a length of feature vectors, one for each location within the (W, H) region. Specifically, in the case of AMNet the feature tensor has dimensions $14 \times 14 \times 1024$. In general there are $L = W \times H$ locations, represented as a vector $\mathbf{x}$:

$$\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_L\} \qquad \mathbf{x}_i \in \mathbb{R}^D \tag{2}$$

All vectors are column vectors, unless stated otherwise. The memorability is estimated with LSTM [9] over a three steps long sequence $T = 3$. The LSTM is defined as:

$$\mathbf{h}_t = \phi(\mathbf{h}_{t-1}, \mathbf{z}_t) \qquad t = [0, T), h \in \mathbb{R}^B \tag{3}$$

where $\mathbf{h}_t$ is the LSTM state at time t with size $B = 1024$. The vector $\mathbf{z}_t$ represents a new image features produced at the step t as a result of the application of the attention weights $\boldsymbol{\alpha}^t$ on the input image features $\mathbf{x}$ and is calculated as a simple weighted sum such that

$$\mathbf{z}_t = \sum_{i=1}^{L} \alpha_{t,i} \mathbf{x}_i \qquad \mathbf{z}_t \in \mathbb{R}^D \tag{4}$$

where $\boldsymbol{\alpha}$ are the attention probabilities conditioned on the entire image feature vector $\mathbf{x}$ and previous LSTM hidden state $\mathbf{h}_{t-1}$

$$\boldsymbol{\alpha}_t \sim p(\boldsymbol{\alpha}_t | \mathbf{x}, \mathbf{h}_{t-1}) \qquad \boldsymbol{\alpha}_t \in \mathbb{R}^L \tag{5}$$

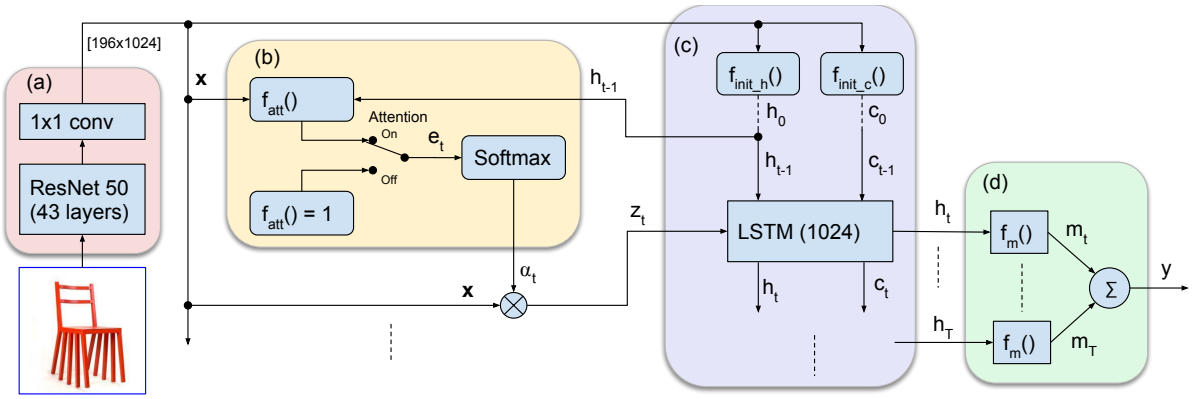

Figure 2: A pretrained RestNet50 (a) is followed by the soft attention mechanism (b) with LSTM (c), which over a sequence of three steps $T = 3$ produces attention maps, each conditioned on the previous LSTM state $\mathbf{h}_{t-1}$ and the entire image feature vector $\mathbf{x}$. Memorability y is then calculated as a sum of discrete memorability scores in the regression network (d).

The attention probabilities, as well as other functions are parameterised with neural networks. The attention is then represented as a vector of weights produced by a softmax function

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{k=1}^{L} \exp(e_{t,k})} \tag{6}$$

The attention weights vector $\mathbf{e}_t$ is a product of the image feature vector $\mathbf{x}$ and the LSTM hidden state $\mathbf{h}_{t-1}$

$$e_{t,i} = f_{att}(\mathbf{x}_i, \mathbf{h}_{t-1}) \tag{7}$$

$f_{att}()$ is s simple sum of two affine transformations followed by logistic function

$$f_{att}(\mathbf{x}_i, \mathbf{h}_{t-1}) = \boldsymbol{M}_i \tanh(\boldsymbol{U}\mathbf{h}_{t-1} + \boldsymbol{K}\mathbf{x}_i + \mathbf{b}) \tag{8}$$

where $\boldsymbol{M}_{L\times D}, \boldsymbol{U}_{D\times B}, \boldsymbol{K}_{D\times D}$ and $\mathbf{b}_{D\times 1}$ are network weights and biases respectively, estimated together with other parameters of the network during optimization.

In order to experiment with the effects of the attention we can conditionally disable it by defining the $f_{att}()$ as a constant function with unit output such that:

$$f_{att}(\mathbf{x}_i, \mathbf{h}_{t-1}) = 1 \tag{9}$$

The results it that all feature vectors in $\mathbf{x}$ are considered equally, thus disabling the attention mechanism.

At each step t the network produces one discrete memorability score m_t calculated as:

$$m_t = f_m(\mathbf{h}_t) \tag{10}$$

The function $f_m()$ maps the LSTM hidden state $\mathbf{h}_t$ to the memorability score $m_t = [0, 1]$. It is implemented as a two-layer neural network for regression with a single output neuron and linear activation function. Finally, the total image memorability score y is calculated as a sum of the discrete memorabilities m_t

$$y = \sum_{t}^{T} m_t \tag{11}$$

In the first step, the LSTM hidden $\mathbf{h}_0$ and memory $\mathbf{c}_0$ states are initialized from the image feature vector $\mathbf{x}$ as follows:

$$\mathbf{c}_0 = f_{init_c}\left(\frac{1}{L}\sum_{i}^{L}\mathbf{x}_i\right) \quad \mathbf{h}_0 = f_{init_h}\left(\frac{1}{L}\sum_{i}^{L}\mathbf{x}_i\right) \tag{12}$$

where the $f_{init}()$ functions are single, fully connected neural networks with $\tanh()$ activation.

3.4. Training Procedure

The AMNet model is trained by minimizing the following loss function:

$$\mathcal{L} = (\hat{y} - y)^2 + \lambda\mathcal{L}_\alpha \tag{13}$$

The first term represents a mean squared error between the ground truth $\hat{y}$ and predicted image memorability y. In order to encourage the attention model to explore all image regions over all time steps, we add a second term $\lambda\mathcal{L}_{\boldsymbol{\alpha}}$ which performs a joint ℓ_1-ℓ_2 penalty as a function of activations of all attention maps in the LSTM sequence T, introduced by Xu et al. [36]. The hyper-parameter λ specifies the impact of this penalty.

$$\mathcal{L}_{s\alpha} = \sum_{i}^{L} s_i^2 \tag{14}$$

s_i represents the ℓ_1 penalty, which enforces sparsity along the sequence dimension T. In other words, it encourages a

strong activation for only one of the attention maps at location i.

$$s_i = 1 - \sum_{t}^{T} \alpha_{t,i} \tag{15}$$

Finally, the ℓ_2 penalty in the form of $\sum_i s_i^2$ in Eq. 14 further promotes an even distribution of activations over all locations. The value of the λ parameter was experimentally determined as 10^{-4} for which the network achieved the highest performance.

The entire model if fully differentiable and trained end-to-end with the ADAM [20] optimizer with a fixed learning rate 10^{-3}. The input image feature vector $\mathbf{x}$ is extracted from the 43^{rd} layer of the RestNet50 [10] with dimensions $[14 \times 14 \times 1024]$. The ResNet50 is trained for image classification on the ImageNet dataset and its weights are not updated during the AMNet training.

The AMNet network is heavily regularized with dropout and with small ℓ_2 weights regularization 10^{-6}. We found that the dropout was critical to stop the network from overfitting. The training was carried out in minibatches of 256 images and terminated by early stopping when the observed Spearman's rank correlation on the validation dataset reached its maximum, which was between epoch 30 and 50 depending on the split and the training dataset (LaMem or SUN). Training and validation losses as well as the memorability rank correlation on the validation dataset in the LaMem, split 1 is shown in Figure 3.

3.5. Data Preprocessing and Augmentation

Common augmentation techniques are applied to the images during the training stage to reduce overfitting and improve generalization. A crop of random size of (0.08 to 1.0) of the original size and a random aspect ratio of 3/4 to 4/3 of the original aspect ratio is made and then resized to 224×224 and randomly, horizontally flipped. For the evaluation only a center crop 224×224 was selected for the input.

Memorability scores in the LaMem dataset are in the range $[0, 1]$ with distribution shown in Figure 4. For the training purpose the memorability scores were zero mean centered and scaled to range $[-1, 1]$.

4. Experimental Results

In this sections we evaluate the AMNet on the LaMem [18] and SUN Memorability [13] datasets. First we briefly describe the datasets and used evaluation metrics, and then present our qualitative and quantitative results with the comparison against the state of the art.

4.1. Datasets

Main focus of this research work is on the LaMem [18] dataset due to its large size which makes it suitable for training deep neural networks. The LaMem is the largest annotated image memorability dataset to this date with total of 58741 images. The images cover a wide range of indoor and outdoor environments, objects and people and were obtained from other labeled datasets such as MIR Flicker, AVA dataset [25], affective images dataset [16], image saliency datasets [23], [29], SUN [35], image popularity dataset [17], Abnormal Objects dataset [31] and a Pascal dataset [8]. The memorability scores were collected manually on the Amazon Mechanical Turk (AMT) by means of a memorability game introduced by [13] and improved by [18]. Approximately 80 measurements (memorable=yes/no) were collected per image. There are 5 random splits each with 45000 images for training, 3741 for evaluation and 10000 for testing.

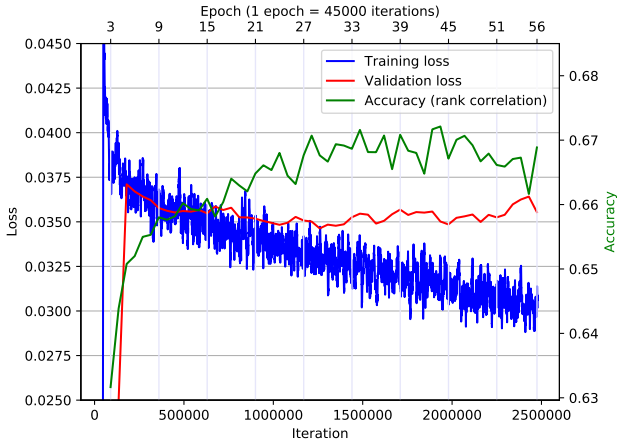

Figure 3: Training/validation losses and memorability rank correlation on the validation dataset in the LaMem split1.

As a second dataset for evaluation we chose the SUN Memorability dataset pioneered by Isola et al. [13]. There are 2222 images in total, originating from the SUN [35] dataset with memorability scores collected similarly to the LaMem. There are 25 random splits with equal number of 1111 images for training and testing.

4.2. Evaluation Metrics

Following the previous work, we report on the performance in terms of rank correlation, specifically a Spearman's rank correlation coefficient [28] ρ and mean squared error MSE.

The Spearman's rank correlation coefficient measures consistency between the predicted and ground truth ranking, within the range $[-1, +1]$ where zero represents no correlation. Higher ρ values indicate better memorability prediction method:

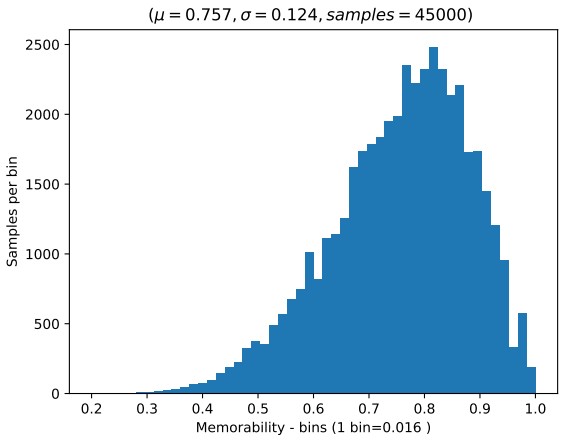

Figure 4: Histogram of ground truth memorability scores in the LaMem [18] training dataset split1.

$$\rho_s(\hat{r}, r) = 1 - \frac{6\sum_i^N (\hat{r_i} - r_i)^2}{N(N^2-1)} \tag{16}$$

where N is a number of samples, $\hat{r_i}$ is a rank of the i^{th} ground truth memorability score, and r_i the i^{th} prediction.

MSE is used as a secondary metric, not always presented in previous work. The Spearman's rank correlation shows a monotonic relationships between the reference and observations but does not reflect the absolute numerical errors between them, which is then presented by MSE according to:

$$\text{MSE}(\hat{y}, y) = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{17}$$

where $\hat{y}_i$ is the ground truth memorability score, while y_i the prediction and N number of tested samples.

4.3. Performance Evaluation

In order to obtain results that are fully comparable with the previous work, we used the same training and evaluation protocol as in the [18] for the LaMem dataset and [13] for the SUN memorability dataset.

Evaluation on the LaMem dataset was performed by training one model on each of the five random splits as suggested by the authors [18] and then reporting the final memorability rank correlation and MSE , averaged over the results from five corresponding test datasets.

In Table 1 we show that the AMNet model with the active attention achieves $\rho = 0.677$, or a 5.8% improvement over the best known method MemNet [18]. Even without attention the AMNet outperforms prior work by 3.6% which demonstrates that the pretrained, deep CNN with our recurrent and regression network layers still achieve high accuracy. The comparatively low performance of the CNN-MTLES [14] method can be attributed to the fact that this model uses various, specifically engineered visual features and features extracted from CNN networks trained on ImageNet [30] and Places [39]. Thus it does not leverage the end-to-end deep learning. The CNN-MTLES, however, uses the LaMem dataset, which indicates that even a large dataset does not significantly improve the performance of models based on engineered visual features.

Method (LaMem dataset)	$\rho \uparrow$	MSE $\downarrow$
AMNet	**0.677**	**0.0082**
AMNet (no attention)	0.663	0.0085
MemNet [18]	0.64	NA
CNN-MTLES [14] (different train/test (50/50) split)	0.5025	NA

Table 1: Average Spearman's rank correlation ρ and MSE over 5 test splits of the LaMem dataset.

Method (SUN Memorability dataset)	$\rho \uparrow$	MSE $\downarrow$
Isola [13]	0.462	0.017
Mancas & Le Meur [24]	0.479	NA
AMNet	**0.649**	**0.011**
AMNet (no attention)	0.62	0.012
MemNet [18]	0.63	NA
MemoNet 30k [2]	0.636	0.012
Hybrid-CNN+SVR [37]	0.6202	0.013

Table 2: Evaluation on the SUN Memorability dataset. All models were trained and tested on the 25 train/val splits.

To train the deep AMNet model on the rather small SUN dataset we had to increase regularization to avoid overfitting. We found that in this specific case $\ell_2 = 10^{-4}$ weights regularization performed better than a stronger dropout or the combination of both. Table 2 shows that the AMNet with attention performs 2% better than the current best model. By disabling the attention the performance declined to $\rho = 0.62$, demonstrating the advantages of visual attention for this task.

We found that during training MSE on the validation datasets follows a similar trend with the rank correlation ρ, however the ρ peaks after the model starts overfitting as seen in Figure 3. It is conceivable to assume that the slightly higher variance at the maximum ρ improves generalization in terms of the predicted and ground truth monotonic relationships, even though MSE starts increasing. For example, during the training on the LaMem split 1, as seen in Figure 3, we attained maximum $\rho = 0.6721$ and MSE $= 0.00848$ while $\rho = 0.6676$ for minimum MSE $= 0.00844$.

Tables 1 and 2 show that the AMNet exhibits the best performance in terms of the Spearman's rank correlation as well as MSE on both, the LaMem and the SUN datasets. The best performance attains $\rho = 0.677$ on the LaMem dataset, approaching 99.6% of the human performance $\rho = 0.68$ as measured by Khosla et al. [18]. Comparison against the state of the art can be seeing in Figure 7.

4.4. The Role of Attention on Memorability

The significant performance gain is achieved by the fact that the neural network learns to focus its attention to specific regions most relevant to memorability. The improvement is close to 2% on the LaMem and almost 5% on the SUN dataset. AMNet learns to explore the image content by producing three visual attention maps, each conditioned on the image content obtained by exploiting the previous map. We have experimented with 2,3,4,5 and 6 LSTM steps and found that three steps are sufficient to achieve the reported performance.

In order to interpret the relation between the attention maps and corresponding discrete memorability estimations in each LSTM step, we converted the attention maps to heat maps and visualized them along with the memorability scores. In Figure 5 we show selected images from the LaMem, split 2 test dataset. Images *(a)*, *(b)* and *(c)* have low memorability, image *(d)* a medium one and *(e)* and *(f)* high memorability. Images of the attention maps are obtained by taking the output of the softmax function Eq. 6, scaled to range $[0, 255]$ and resized from 14×14 to 244×244.

As we can see in images *(a)*, *(c)* and *(d)* in Figure 5, most of the first attention weights gravitate towards the image center, which is most likely caused by the Center Bias, studied in [16], [38] and attributed primarily to the photographer bias. In the subsequent LSTM steps, however, the attention usually moves to the regions responsible for memorability.

After a close inspection, we found that the attention maps for low memorability images tend to be sparser with few small peaks, while for higher image memorability, the attention maps display sharper focus covering larger regions around the activation peaks. Core image memorability usually originates in regions with people and human faces as evident in images *(c)* and *(f)* in Figure 5.

Moreover, we found that the estimates of discrete memorabilities m_t in Eq. 10 decrease with each LSTM step t for low memorability images, while for high memorability images they grow. This relation is shown in Figure 6. This effects is consistent within the LaMem test datasets across all splits and can be seen in Figure 5.

Initially, we experimented with additional penalty function that would encourage the optimizer to estimate the discrete memorabilities in ascending or descending order, however this always caused a drop in the performance. The above observation explains this effect, that is, the gradient of the discrete memorabilities over the LSTM steps differs depending on the core image memorability. Thus forcing the optimizer to maintain positive or negative gradient has a detrimental effect on the model convergence.

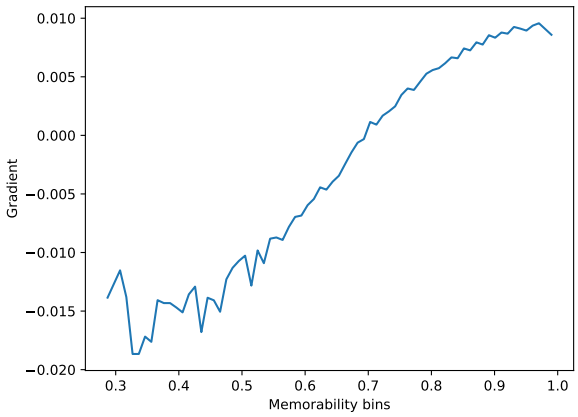

Figure 6: Histogram of gradients of discrete memorabilities over the LSTM steps. The gradient is directly proportional to the total image memorability.

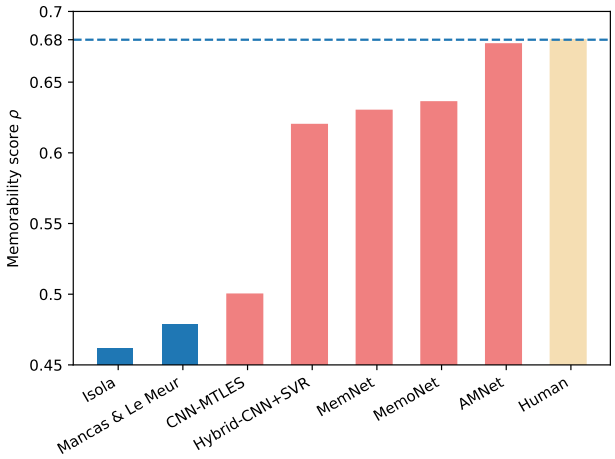

Figure 7: Comparison against the state of the art methods. Red depicts deep learning based methods. AMNet, MemNet and CNN-MTLES [14] where trained on the LaMem, the rest on the SUN Memorability dataset.

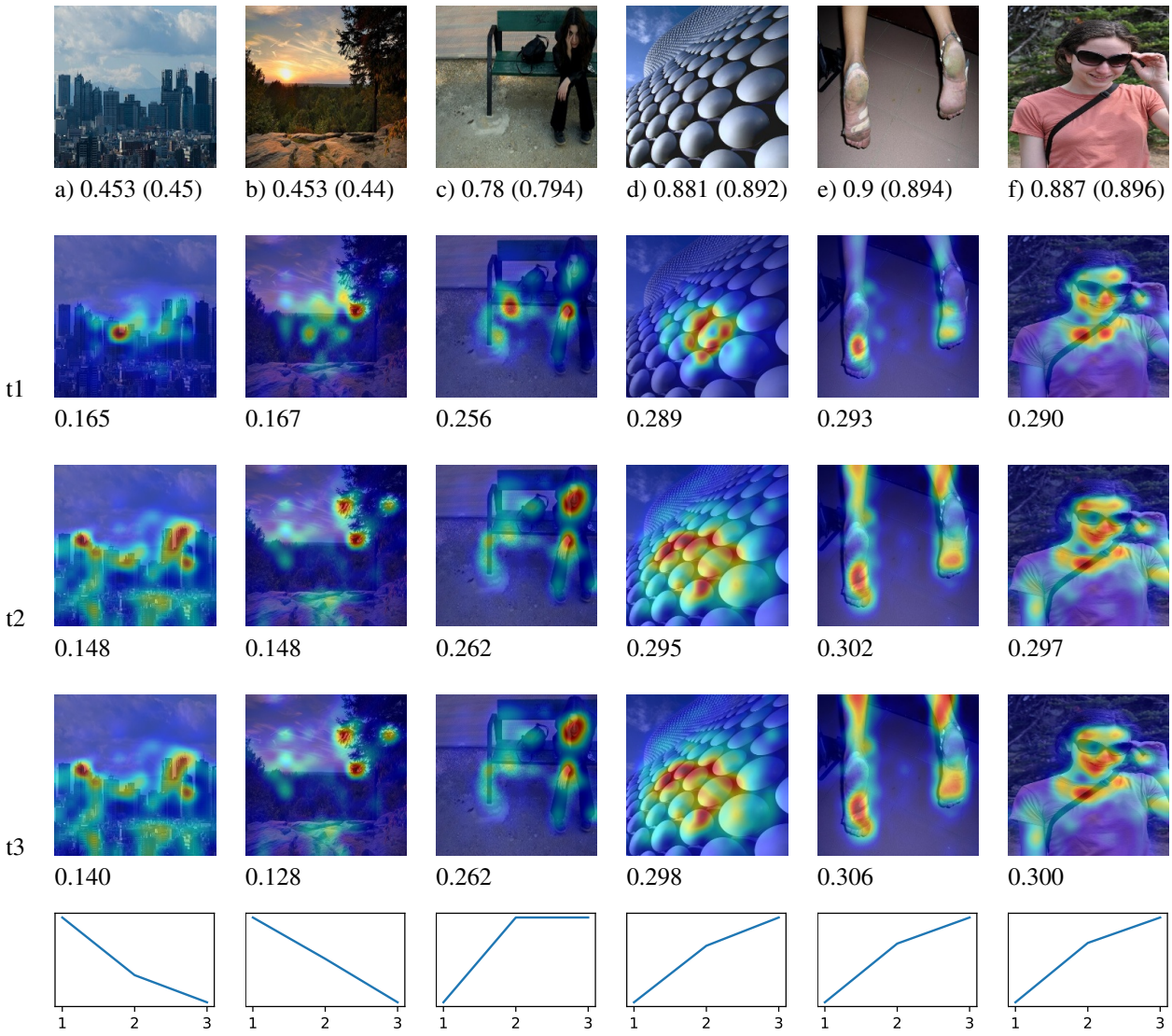

Figure 5: Examples of attention maps for low and high memorability images from LaMem test dataset split 2. Tested images, their estimated and ground truth memorabilities (in brackets) are shown in the top raw. Bellow each image is a discrete memorability score estimated at the steps t_1, t_2 and t_3. Plots at the bottom row show gradients over three LSTM steps.

5. Conclusion

In this work we propose AMNet, a novel deep neural network with visual attention component for image memorability estimation. This network consists of a pre-trained, deep CNN followed by a modified visual attention mechanism with a recurrent network and network for memorability regression. By design the AMNet is generic and could be employed for other regression, computer vision tasks.

We show that a deep CNN, trained on large-scale image classification is beneficial for the memorability estimation task, indicating that the feature hierarchies extracted for the image classification are suitable to express the composition underlying the memorability effect.

Finally, we demonstrate that our recurrent visual attention network significantly improves performance of the image memorability learning and inference.

The proposed method outperforms previous state of the art work by 5.8% (from $\rho = 0.64$ to $\rho = 0.677$) on the Spearman's rank correlation and closely approaches the human performance $\rho = 0.68$ with a 99.6% consistency. The AMNet implementation in PyTorch is available at `https://github.com/ok1zjf/amnet/`

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[2] Y. Baveye, R. Cohendet, M. Perreira Da Silva, and P. Le Callet. Deep learning for image memorability prediction: The emotional bias. In *Proceedings of the ACM International Conference on Multimedia*, pages 491–495, New York, NY, USA, 2016. ACM.

[3] Z. Bylinskii, P. Isola, C. Bainbridge, A. Torralba, and A. Oliva. Intrinsic and extrinsic effects on image memorability. *Vision Research*, 116(Part B):165–178, 2015. Computational Models of Visual Attention.

[4] B. Celikkale, A. Erdem, and E. Erdem. Visual attention-driven spatial pooling for image memorability. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 976–983, 2013.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the Computer Vision and Pattern Recognition*, volume 1, pages 886–893. IEEE, 2005.

[6] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik. Support vector regression machines. In *Advances in Neural Information Processing Systems 9*, pages 155–161. MIT Press, 1997.

[7] R. Dubey, J. Peterson, A. Khosla, M.-H. Yang, and B. Ghanem. What makes an object memorable? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1089–1097, 2015.

[8] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1785. IEEE, 2009.

[9] K. Greff, R. K. Srivastava, J. Koutn'ik, B. R. Steunebrink, and J. Schmidhuber. LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 2017.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[11] P. Isola, D. Parikh, A. Torralba, and A. Oliva. Understanding the intrinsic memorability of images. In *Advances in Neural Information Processing Systems*, pages 2429–2437, 2011.

[12] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva. What makes a photograph memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1469–1482, July 2014.

[13] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 145–152. IEEE, 2011.

[14] P. Jing, Y. Su, L. Nie, and H. Gu. Predicting image memorability through adaptive transfer learning from external sources. *IEEE Transactions on Multimedia*, 19(5):1050–1062, 2017.

[15] T. Joachims. Training linear SVMs in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 217–226. ACM, 2006.

[16] Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *International Conference on Computer Vision (ICCV)*, pages 2106–2113, 2009.

[17] A. Khosla, A. Das Sarma, and R. Hamid. What makes an image popular? In *Proceedings of the international conference on World wide Web*, pages 867–876. ACM, 2014.

[18] A. Khosla, A. S. Raju, A. Torralba, and A. Oliva. Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2390–2398, 2015.

[19] A. Khosla, J. Xiao, A. Torralba, and A. Oliva. Memorability of image regions. In *Advances in Neural Information Processing Systems*, pages 296–304, 2012.

[20] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

[22] D. G. Lowe. Distinctive image features from Scale-Invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[23] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 83–92. ACM, 2010.

[24] M. Mancas and O. Le Meur. Memorability of natural scenes: The role of attention. In *International Conference on Image Processing*, pages 196–200. IEEE, 2013.

[25] N. Murray, L. Marchesotti, and F. Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *Computer Vision and Pattern Recognition*, pages 2408–2415. IEEE, 2012.

[26] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[27] H. Peng, K. Li, B. Li, H. Ling, W. Xiong, and W. Hu. Predicting image memorability by multi-view adaptive regression. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1147–1150. ACM, 2015.

[28] W. Pirie. Spearman rank correlation coefficient. *Encyclopedia of statistical sciences*, 1988.

[29] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.-S. Chua. An eye fixation database for saliency detection in images. *European Conference on Computer Vision*, pages 30–43, 2010.

[30] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, and Others. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[31] B. Saleh, A. Farhadi, and A. Elgammal. Object-centric anomaly detection by attribute-based reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 787–794, 2013.

[32] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*, pages 806–813, 2014.

[33] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9. IEEE, 2015.

[35] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3485–3492. IEEE, 2010.

[36] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.

[37] S. Zarezadeh, M. Rezaeian, and M. T. Sadeghi. Image memorability prediction using deep features. In *Iranian Conference on Electrical Engineering (ICEE)*, pages 2176–2181. IEEE, 2017.

[38] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. SUN: A bayesian framework for saliency using natural statistics. *Journal of vision*, 8(7):32, 2008.

[39] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, pages 487–495, Cambridge, MA, USA, 2014. MIT Press.